New Evidence of Factor Structure and Measurement Invariance of the SDQ Across Five European Nations

Javier Ortuño-Sierra[1], Eduardo Fonseca-Pedrero[1], Rebeca Aritio-Solana[1], Alvaro Moreno Velasco[1], Edurne Chocarro de Luis[1], Gunter Schumann[2,3]*, Anna Cattrell[2,3], Herta Flor[4], Frauke Nees[4], Tobias Banaschewski[5], Arun Bokde[6], Rob Whelan[6],Christian Buechel[7], Uli Bromberg[7], Patricia Conrod[2,8], Vincent Frouin[9], Dimitri Papadopoulos[9], Juergen Gallinat[10], Hugh Garavan[11], Andreas Heinz[10], Henrik Walter[10], Maren Struve[12], Penny Gowland[13], Tomáš Paus[14], Luise Poustka[5], Jean-Luc Martinot[15], Marie-Laure PaillèreMartinot[15], Nora C. Vetter [16], Michael N. Smolka[17], Claire Lawrence[17], and the IMAGEN consortium.

[1]Department of Educational Sciences, University of La Rioja, Spain; [2] Institute of Psychiatry, Psychology & Neuroscience, King's College London, United Kingdom; [3]Medical Research Council – Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, United Kingdom, De Crespigny Park, London, United Kingdom;[4]Department of Cognitive and Clinical Neuroscience, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Square J5, Mannheim, Germany; [5]Department of Child and Adolescent Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Square J5, 68159 Mannheim, Germany; [6]Discipline of Psychiatry, School of Medicine and Trinity College Institute of Neurosciences, Trinity College Dublin; [7]University Medical Centre Hamburg-Eppendorf, Haus S10, Martinistr. 52, Hamburg, Germany; [8]Department of Psychiatry, Universite de Montreal, CHU Ste Justine Hospital, Canada; [9]Neurospin, Commissariat à l'EnergieAtomique, CEA-Saclay Center, Paris, France;[10]Department of Psychiatry and Psychotherapy, Campus CharitéMitte, Charité,Universitätsmedizin Berlin, Charitéplatz 1, Berlin, Germany; [11]Departments of Psychiatry and Psychology, University of Vermont, 05405 Burlington, Vermont, USA; [12]Praxis für Psychologische Psychotherapie, Ilvesheimer Str. 17, 68259 Mannheim[12]Physikalisch-Technische Bundesanstalt, Abbestr. 2-12, Berlin, Germany; [13]School of Physics and Astronomy, University of Nottingham, United Kingdom; [14]The Rotman Research Institute, University of Toronto;[15]Institut National de la Sante et de la Recherche Medicale, INSERM CEAUnit1000, ''Imaging & Psychiatry'', University Paris Sud, 91400 Orsay, France; [16]Department of Psychiatry and Neuroimaging Center, Technische Universität Dresden, Dresden, Germany; [17]School of Psychology, University of Nottingham, United Kingdom

---

Corresponding Author:
Javier Ortuño Sierra
Department of Educational Sciences
University of La Rioja
C/ Luis de Ulloa, s/n, Edificio VIVES
PC: 26002, Logroño, La Rioja, Spain
Telephone: +34 941 299 309
Fax: +34 941 299 333
E-mail: javier.ortuno@unirioja.es

Acknowledgements: This work received support from the following sources: the European Union-funded FP6 Integrated Project IMAGEN (Reinforcement-related behaviour in normal brain function and psychopathology) (LSHM-CT- 2007-037286), the FP7 projects IMAGEMEND(602450; IMAgingGEnetics for MENtal Disorders) and MATRICS (603016), the Innovative Medicine Initiative Project EU-AIMS (115300-2), a Medical Research Council Programme Grant ‘‘Developmental pathways into adolescent substance abuse’’ (93558), the Swedish funding agency FORMAS, the Medical Research Council and the Wellcome Trust (Behavioural and Clinical Neuroscience Institute, University of Cambridge), the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London, the Bundes ministerium für Bildung und Forschung (BMBF grants 01GS08152; 01EV0711; eMEDSysAlc01ZX1311A; Forschungsnetze AERIAL and BipoLife) and the Deutsche Forschungsgemeinschaft (DFG grants FOR 1617, SFB 940 & SM 80/5-2).

Abstract

The main purpose of the present study was to test the internal structure and to study the measurement invariance of the Strengths and Difficulties Questionnaire (SDQ), self-reported version, in five European countries. The sample consisted of 3012 adolescents aged between 12 and 17 years (M = 14.20; SD = .83). The five-factor model (with correlated errors added), and the five-factor model (with correlated errors added) with the reverse-worded items allowed to cross-load on the Prosocial subscale, displayed adequate goodness of-fit indices. Multi-group confirmatory factor analysis showed that the five-factor model had partial strong measurement invariance by countries. A total of 11 of the 25 items were non-invariant across samples. The level of internal consistency of the Total difficulties score was .84, ranging between .69 and .78 for the SDQ subscales. The findings indicate that the SDQ's scales need to be modified in various ways for screening emotional and behavioural problems in the five European countries that were analyzed.

Keywords: SDQ; Self-report; Adolescents; Factorial Structure; Measurement Invariance; Behavioural problems

Adolescence is a particularly important developmental stage for socio-emotional development, but it is also marked by the emergence of mental health problems [1]. These problems are both common and debilitating during adolescence producing significant social and economic consequences for the individual, their families, and the global community [2,3]. As a consequence, interest in the detection of children and adolescents at risk for emotional disorders or behavioural problems have sharply increased in the last two decades [4-7]. The assessment of emotional and behavioural problems in children and adolescents is a priority issue not only for public health policy, but also in the context of clinical practice and research. Nevertheless, and despite the efforts in early detection, different research studies have suggested that only a minority of the adolescent population with needs in the area of mental health comes in direct contact with specialized services [8,9].

Among the measuring instruments developed to assess psychological difficulties and capacities in children and adolescents we can find the Strengths and Difficulties Questionnaire (SDQ) [10] self-report version. The SDQ is a screening instrument for behavioural and emotional problems that also assesses capacities in the social sphere. Furthermore, it is a brief, simple, and easy management tool for use in child and adolescent populations [11,12]. It has been used in both clinical and community settings throughout the world. The SDQ is composed of 25 items in a Likert response format with three response options grouped into five subscales [10]: Emotional symptoms, Conduct problems, Hyperactivity, Peer problems, and Prosocial behaviour. The first four subscales form a Total difficulties score. Items that compose the SDQ are both positively and negatively phrased in order to avoid the effect of response bias (e.g., acquiescence). In total, 15 items reflect problems and 10 capabilities, of which five belong to the Prosocial subscale and five should be recoded, since they are formulated in a positive way and belong to the problems subscales.

Previous studies have reported adequate psychometric properties related to reliability and sources of validity evidences for the SDQ self-reported version [13-15]. Nevertheless, several studies have detected low values of reliability (Cronbachs's alpha < .60), especially in the Conduct problems and Peer problems subscales [16-23,11,24,25]. An added value of the test, as it the fact that it contains positive items, could be a key factor in explaining low internal consistency and the inconsistency of factorial solutions [26]. The fact that the problems subscales includes this type of items can generate that they behave as part of a distinct construct [20]. Therefore, reverse-worded items may influence the estimation of internal consistency due to their low correlation with the rest of the SDQ items that

measure problems, and could, at the same time, affect the factor structure [26]. In this sense, a five-factor solution of the SDQ with the same 25 items but allowing the reverse-worded items to cross-load on the Prosocial subscale has been proposed as the most satisfactory [26].

With regards to the study of the factor structure of the SDQ, there are contradictory results. Previous studies, conducted using confirmatory factor analysis, showed the five-factor solution as the most appropriate [27,24,11,28,29,25], while others found the three-factor model [30-32,11]. The three-factor solution is composed by Internalizing symptoms, resulting of the Emotional and Peer problems subscales, the Externalizing symptoms, composed by Conduct problems and Hyperactivity subscales, and the Prosocial subscale. In addition other studies have found a five-factor solution with two second order factor (internalizing and externalizing) as the most satisfactory model [33]. Nevertheless, Mellor and Stokes [22] reported that none of the five subscales was essentially one-dimensional, questioning the adequacy of the internal structure of the five-factor solution. Other research, likewise, has discussed the adequacy of the of subscales, concluding that the SDQ internal structure was not appropriate [19,32]. Also, a bifactor structure of the test has been also found to be adequate [34]

Another important issue regarding the factor structure of the SDQ is the study of Measurement Invariance (MI) across nations. Recently, Goodman et al. [35] suggested, according to their results, that cross-national differences in SDQ scores may be due to different biases instead of reflecting comparable differences in mental disorder rates. Thus, cross-national differences in SDQ caseness may be determined by its measuring construct (i.e., factorial structure) rather than real comparable differences in disorder rates. In the same line, Milfont and Fisher [36] pointed out that MI has to be demonstrated for a meaningful comparison of measuring constructs across groups. The evaluation of MI is important for determining the generalizability of latent constructs across groups and whether the MI and the construct being measured are operating in the same way across diverse samples of interest [37]. If MI does not hold, inferences and interpretations drawn from the data may be erroneous or unfounded. It is also a priority to conduct studies of measurement equivalence that guarantee the comparability of scores across cultures (e.g., to set cut-off scores, to conduct cross-cultural comparisons).

Previous studies have analyzed the MI of the SDQ, self-reported version in adolescents, across different variables (e.g., gender, age, race/ethnicity, and income) [19,11,26,38-40]. As yet, there has been no in-depth examination addressing the question of whether the dimensional structure underlying

the SDQ scores is invariant across countries. To the best of our knowledge, few studies have addressed this question [39,38,27]. For instance, Essau et al. [38], with a sample of 2418 adolescents, found that the factorial structure of the SDQ differed across five European countries (Cyprus, England, Sweden, Germany, and Italy). Another important study, conducted by Stevanovic et al. [39], did not find an acceptable model in countries from Europe, Asia, and Africa (India, Nigeria, Turkey, Croatia, Indonesia, Bulgaria, and Serbia). As so, they were not able to test for MI across countries.

Thus, the replicability of the factorial structure of the SDQ in its self-report form with adolescent population across different cultural groups is a question that still needs further examination. Within this framework, the main goal of the present study was to study the internal structure of the SDQ scores and to test the equivalence of the factor structure of the SDQ across five European countries. We therefore intended to study the internal structure, we tested the measurement invariance across different countries, and we studied the internal consistency of the SDQ scores using Ordinal alpha. We hypothesized that a five-factor model allowing reverse-worded items to cross-load on the Prosocial factor would provide the best fit to the data in all the countries. We further hypothesized that the factor structure underlying the SDQ scores would be invariant across cultures. Moreover, we hypothesized that internal consistency of the scores would be adequate for the Total difficulties score and in all of the subscales, with possible decrease in the Conduct and Peers problems subscales according to previous studies [16-23,11,24,25].

Method

Participants

A total of 3260 students completed the SDQ questionnaires. Cases with missing data on gender, age, and SDQ items were excluded. Final sample comprised a total of 3012 adolescents, 1434 were male (47.6%), from five European countries: Spain (N = 848; 28.2%), England (N = 626; 20.8%), Ireland (N = 227; 7.5%), Germany (N = 1050; 34.9%), and France (N = 260; 8.6%). Participants' ages ranged between 12 and 17 years (M = 14.20; SD = 0.83). In all samples, students were from different types of secondary schools – public, grant-assisted private and private – and from vocational/technical schools.

The age distribution of the sample in Spain was the following: 12 years (n = 70; 8.2%), 13 years (n = 107; 12.6%), 14 years (n = 181; 21.3%), 15 years (n = 222; 26.1%), 16 years (n = 165;

19.4%), and 17 years (n = 104; 12.2%). All of the adolescents in England, Ireland, Germany, and France were 14 years old. As a consequence, statistically significant differences were found by age ($F_{(5,1910)} = 117.02$, $p \leq .001$).

With regards to the gender, the distribution of the total sample was the following: Spain (male = 368; 43.3%), England (male = 310; 49.5%), Ireland (male = 125; 55.1%), Germany (male = 499; 47.5%), and France (male = 132; 50.8%). No statistically significant differences were found by gender ($F_{(4,748)} = 3.31$, $p > .001$) across the countries.

Instrument

The Strengths and Difficulties Questionnaire (SDQ) [10], self-reported form. It is a measuring instrument widely used for the assessment of different social, emotional and behavioural problems related to mental health in children and adolescents over the previous six months. The SDQ is made up of a total of 25 statements distributed across five subscales (each with five items): Emotional symptoms, Conduct problems, Hyperactivity, Peer problems, and Prosocial behaviour. In this study we used a Likert-type response format with three options (0 = "*Not true*", 1 = "*Somewhat true*", 2 = "*Certainly true*"), so that the score on each subscale ranged from 0 to 10 points.

In the present study, for the Spanish sample we used the version Spanish validated in non-clinical adolescent populations [40,41]. The original English version of the manuscript was used in the case of the English and the Irish samples [10]. The validated German version of the SDQ [14] was used in the case of Germany. Finally the validated French version of the instrument [16] was used in the French sample.

Procedure

In Spain, the questionnaire was administered collectively, in groups of 10 to 35 students, during normal school hours and in a classroom specially prepared for this purpose. In the remaining countries, data collection took place as part of a larger study examining adolescent reinforcement-related behaviour in non-clinical populations. The SDQ was completed individually using a computer-based system at research institutes in Nottingham, London, Paris, Berlin, Mannheim, Dresden, Hamburg and Dublin. For details of the larger study please refer to Schumann et al. [42]. School approval and parental written informed consent were obtained in all the countries for participation. Participants were informed of the confidentiality of their responses and of the voluntary nature of the

study. No incentives were given for completing the SDQ specifically. Administration took place under the supervision of researchers.

Data analyses

First, in order to analyse the internal structure of SDQ scores and based on previous literature, several confirmatory factor analyses (CFAs) were conducted at the item level. Due to the categorical nature of the data, we used the robust Mean-adjusted Weighted Least Square method (WLSMV) for the estimation of parameters [43]. The following goodness-of-fit indices were used: Chi-square (χ^2), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (WRMR). The CFI and TLI values greater than .95 are preferred and values close to .90 are considered acceptable, and WRMR values less than .08 are considered a proof of good model fit, while the RMSEA values should be under .08 for a reasonable fit, and under .05 for a good fit [44,45].

Based on previous literature, different hypothetical dimensional: a) the original five-factor model [10], b) the five-factor model (with five correlated errors allowed), c) a five-factor model with correlated errors and the reverse-worded items allowed to cross-load on the Prosocial factor [26], d) the three-factor model which is composed by Internalizing symptoms, resulting of the Emotional and Peer problems subscales, the Externalizing symptoms, composed by Conduct problems and Hyperactivity subscales, and the Prosocial subscale [31], e) a three-factor model with correlated errors [26], f) a three-factor model with correlated errors and allowing reverse-worded items to cross-load on the Prosocial factor, g) the five-factor model with two second-order factors [33], resulting from grouping internalizing symptoms (emotional and peer subscales) and externalizing symptoms (behavioural and hyperactive subscales) and the Prosocial factor, h) the five-factor model with two second-order factors and correlated errors, and i) the five-factor model with two second-order factors with correlated errors and allowing reverse-worded items to cross-load on the Prosocial factor.

Since some correlation errors were found in the original five-factor model, we decided to allow the correlation between those items that had similar content: 2-10 (I am restless-constantly fidgeting), 25-15 (finish the work-easily distracted), 15-16 (easily distracted- nervous in new situations), 19-18 (others bully me- often accused of lying), and 20-23 (volunteer to help others-get on better with adults). Some other correlated errors were identified. However, taking into account the inherent

problematic in the use of correlated errors [46], and from a pragmatic criterion we decided to compute just five correlated errors in the analyses.

Second, we tested MI. Successive multigroup CFAs were conducted [47]. A hierarchical set of steps are followed when MI is tested typically starting with the determination of a well-fitting multigroup baseline (configural) model and continuing with the establishment of successive equivalence constraints in the model parameters across groups. Using Delta parameterization in Mplus, two steps on MI need to be considered: Configural and strong invariance models. As proposed by Muthén and Muthén [48], analyses of measurement invariance with delta parameterization for binary and ordered polytomous data are conducted considering metric and scalar invariance in tandem. Thus, we decided to test for measurement invariance attending to configural and strong invariance models. The configural model is the first and least restrictive model to be tested. The configural model is established by specifying and testing the model for each group separately. Once the theoretical model has been validated in all groups, configural invariance is examined requiring that the same pattern of fixed and freely estimated parameters are equivalent across groups, and therefore, that no equality constraints are imposed. When the configural invariance model is found, it is assumed that the general factor structure is at least similar, though not necessarily equivalent, across groups. In a second step, we established a strong invariance model, which contained cross-group equality constraints on all factor loadings and item thresholds. As required by the model, scale factors were fixed to one in one group and were free in the others, and factor means were fixed to zero in one group and were free in the others [43]. The Spanish group was set as the reference group. The assumption of strong invariance model is also necessary for comparing groups [47,49-51].

The analysed dimensional models can be seen as nested models to which constraints are progressively added. Due the $\Delta\chi^2$sensitivity to sample size, Cheung and Rensvold [52] proposed a more practical criterion: the change in CFI (ΔCFI), to determine if nested models are practically equivalent. In this study, when ΔCFI is greater than .01 between two nested models, the more constrained model is rejected since the additional constraints have produced a practically worse fit. However, when this criterion is not met and some of the parameters (e.g., factorial loadings or thresholds) are not specified to be equal across groups, partial measurement invariance model can be considered [53].

Finally, we calculated internal consistency and descriptive statistics of the SDQ subscales and Total difficulties score for each country and the total sample. Ordinal alpha coefficient for Likert data

was calculated as a measure of the reliability of the SDQ scores. Ordinal alpha is conceptually equivalent to Cronbach's alpha and it is more adequate for dichotomous and ordinal data.SPSS 15.0 [54], Factor 9.2, and Mplus 7.0 [43], were used for data analysis.

Results

Validity evidence based on internal structure of the SDQ scores: confirmatory factor analysis

CFAs showed that the five-factor model (with correlated errors), allowing reverse-worded item to cross-load on the Prosocial factor, displayed better goodness-of-fit indices than the other hypothetical dimensional models tested in Spain, England, Ireland, and in the total sample. Nevertheless, the five-factor model with the inclusion of correlated errors, showed similar results in Germany and France. As it is shown in Table 1, goodness-of-fit indices for the baseline five-factor and three-factor models did not reach the cut-offs recommended. For both models, substantial Modification Indices (MIs) were found for error correlation between items 25 and 15, items 2 and 10, items 19 and 18, items 20 and 23, and items 15 and 16. This correlation was made between those items that have similar content. Once the correlated errors were added, goodness-of-fit indices were adequate for the five-factor solution in all the countries and in the Total sample, with the exception of Ireland (CFI = .842). However, other fit indices, in the case of Ireland, showed adequate indices (RMSEA = .052). It is worth noting that Ireland was the country with the smaller sample size. In this sense, some fit indices, especially when data are considered ordinal and WLSMV is used, can be affected. For this reason, the RMSEA has to be also considered as an even more relevant criterion of fit indices when categorical data are analyzed [55].

When compared, the solution with five factors, displayed better goodness-of-fit than the three-factor solution in all the countries. Meanwhile, the model with the inclusion of second-order factors revealed lower goodness-of-fit indices than the five-factor model in all the countries and in the Total sample.

The standardized factor loadings for the strong partial measurement invariance model for each country are shown in Table 3. All factor loadings were statistically significant in the five countries ranging from .33 (item 11, Ireland) to .95 (item 13 Ireland).

----------Please Insert Table 1 here----------

Measurement invariance of the SDQ scores across countries

Given that the five-factor model with modifications displayed adequate goodness-of-fit indices and as factor loadings and internal consistency levels in this model were more appropriate than model c, we therefore tested the factorial equivalence of this model across countries. The configural invariance model, in which no equality constraints were imposed, showed an adequate fit to the data (see Table 2). Next, a strong invariance model was tested with the item thresholds and factor loadings being constrained to equality across groups. The ΔCFI between the constrained and the unconstrained models was over .010, indicating that strong invariance was not supported. Factor loadings and thresholds of items 2, 15, 20, and 21 in the case of England, items 1, 6, 7, 15, 16, 17, 21, 22, and 23 in Germany, and item 21 in France were freed, meaning that the factor loadings and thresholds of these items were non-equivalent across countries. No items had to be freed in Ireland. In sum, a total of eleven items were non-invariance across the countries. Once the item parameters were freed the model fit was adequate, indicating that strong partial measurement invariance was supported across the countries. The total amount of parameters found common among countries was over 80%.

The standardized factor loadings for the strong partial measurement invariance model for each country are shown in Table 3. All factor loadings were statistically significant in the five countries ranging from .39 (item 10, Spain) to .97 (item 13 Ireland). As shown in Table 3 non-equivalent items belong to all the dimension, with Hyperactivity (2, 15, and 21) and Prosocial (1, 17, and 20) subscales showing a total of three non-equivalent items and Emotional (16) showing just one

----------Please Insert Tables 2 and 3 here----------

Internal consistency and descriptive statistics of the SDQ scores

The internal consistency of the scores by means of Ordinal alpha was calculated (model b). As it is shown in Table 4, Ordinal alpha values for the Total difficulties score (20 items) were good, ranging from .75 (Germany) to .85 (France). Ordinal alpha values in the other subscales (5 items) were also adequate in almost all the countries. Nevertheless, lower values were found in the case of the Conduct problems (.61, France) and the Peer problems (.61, Ireland) subscales.

In addition, descriptive statistics (mean and standard deviation) of the SDQ were calculated for each country and for the total sample (see Table 5).

----------Please Insert Table 4 and 5here----------

Discussion and Conclusions

The main purpose of this study was to analyse the internal structureand the measurement invariance ofthe Strengths Difficulties Questionnaire (SDQ) [10] in its self-reported form using a large sample of adolescents from five European countries. To this end, we examined the internal structure, tested the measurement invariance across countries, and estimated the internal consistency of the SDQ scores. Knowledge of the SDQ psychometric properties is relevant to use it as a screening tool in an age group at particular risk of developing emotional and behavioural symptoms and disorders [1,4-7].

The study of the internal structure, by means of CFAs, supported the five-factor structure in all the countries and in the total sample, as it is the case in previous studies [40,27,11,24,28,29,25]. Nevertheless, adequate goodness-of-fit indices were found after adding error correlation between items, indicating discrete values in the five-factor baseline model in all the countries. Moreover, some goodness-of-fit indices of Ireland were still not appropriate. Similar results were found in previous studies [40,32,19,26]. For instance, the study of Ortuño-Sierra et al. [40] showed that the five-factor structure was the better to fit the data, but appropriate goodness-of-fit were only reached after correlated errors were added. Thus, the five-factor structure is still questionable. In the same line, a modified five factor model allowing the reverse-worded items to cross-load on the Prosocial factor displayed significant better goodness-of-fit indices in all the countries, including the total sample, as it was the case in the study of van de Looij–Jansen et al. [26]. However, the study of factor loadings revealed that some of them were non-significant, questioning the adequacy of this model.

With regards to the three-factor structure, the results of the CFAs indicated lower goodness-of-fit-indices than the five-factor model. The respective models based on the three-factor structure with the correlated errors added, and with the reverse-worded items allowed to cross-load on the Prosocial dimension, displayed all of them lower fit indices than their correlated five-factor models. As so, the three-factor structure of the SDQ was found not adequate, similarly to the findings in previous studies [40,56,32,11]. Regarding this, ΔCFI analysis revealed that, contrary to van de Looij-Jansen et al.[26], both in the three and the five-factor models, the inclusion of modifications is more significant in model fit than allowing reverse-worded items to cross-load on the Prosocial subscale with, the exception of

Ireland. Nonetheless, in all cases, the extension of the Prosocial subscale resulted in an improvement of the model fit in all the countries, confirming the results of van de Looij-Jansen et al.[26], and the idea that the extended Prosocial factor may reflect the possibility of a positive response construct [20]. However, the study of the factor loadings revealed some non-significant factor loadings in this model, and also levels of internal consistency were less adequate. For this reason, we decided that model b was more appropriate to further study measurement invariance.

Adolescence is a developmental stage in which relevant biopsychological changes occur. These changes could be different depending on factors such as the geographical area, the culture, the meaning of the items or the language [57]. For this reason, we believe that the study of the measurement invariance is important in order to assure the comparability of scores and for determining the generalizability of latent constructs across these groups. The detection of psychological difficulties as well as the prosocial capabilities is a key factor that will allow future intervention with adolescents. Nonetheless, the review of the literature shows that there are few studies of measurement invariance in the self-reported version of the SDQ [19,11,27,26,38,39].

Results supported the hypothesis of partial measurement invariance by nation. Nevertheless, it is worth noting that measurement invariance was reached after factor loadings and threshold of items were freed (1, 2, 6, 7, 15, 16, 17, 20, 21, 22, and 23), in the five-factor model with correlated errors. Therefore, according to the results found, these items should be considered carefully when using the SDQ across countries as data shows that the original five-factor structure is not appropriate for cross-cultural comparisons, as well as the other models tested. Thus, and taking into account previous studies and the results found in the present study, it is possible to affirm that the SDQ can not be used in cross –cultural comparisons, when multiple samples are included. This issue does not imply that SDQ cannot be used for in-country assessments. Recent studies have found similar results, indicating that the structure of the SDQ self-reported version in adolescents was non-invariant across cultures [38,39]. For instance, in the study of Essau et al. [38] with five European countries, the factorial equivalence of the SDQ was rejected in the five and the three dimensional models tested.

In this sense, the countries involved in both studies could be a key factor explaining the differences. It might be that countries involved in the study of Essau et al. [38], were more distant among each other than those involved in the present work. As it is the case of our study, they considered central Europe (Germany), Anglo-Saxon (England), and Mediterranean (Cyprus and Italy)

countries, but they also included a Scandinavian country (Sweden). Thus, future studies should consider the possibility of analysing factorial equivalence among different cultures or geographical areas of Europe like the north and the south. Moreover, parameters were estimated with MLM. In this sense, the different estimator used might explain the differences found with the present study. Also, the study of Stevanovic et al. [39] showed that due to the lack of fit, the study of structural equivalence across the nations was not possible. As it was noted before, it might be because countries from different continents were involved: Europe (Turkey, Serbia, Bulgaria, and Croatia), Asia (India and Indonesia), and Africa (Nigeria).

The results found in the present study give some light in the possibility of the comparison of the SDQ scores between different nations in Europe. Nonetheless, it is worth mentioning that the results have to be considered carefully, as strong measurement invariance was supported after different items were released, indicating differential item functioning in a total of 11 items. As it was proposed by Byrne et al. [53], in a situation where there is no perfect type of measurement invariance (i.e. full measurement invariance), but neither is complete non-invariance, it is possible to talk about partial measurement invariance. In the case of partial measurement invariance presence, only those items that meet criteria for strong measurement invariance model should be included in composited measures when scores for the scales are to be compared cross-culturally. Previous studies have found psychological constructs like emotional and behavioural problems to be invariant across cultures with other measurement instruments (e.g. Youth Self Report) [58].

With regard to the study of the internal consistency of the SDQ scores, adequate levels of reliability were found with Ordinal alpha (0.83) for the Total score in the total sample. Ordinal alpha in the countries ranged from .75 (Germany) to .85 (France). Although still adequate, lower internal consistency values were found in the Conduct and Peer problems subscales, similarly to previous studies [40,18,25,11,24,59,22,17,20,19]. It is worth noting that previous studies analyzing the internal consistency of the SDQ were made through Cronbach's Alpha. In this sense, the fact that Ordinal alpha was used might be a relevant variable that explains these differences. Ordinal alpha, has been shown to estimate reliability more accurately than Cronbach's alpha for ordinal response scales [60]. Also, as it has been proposed, the inclusion of positive items in the problems subscales may affect to the internal consistency of these subscales [26]. In addition, it is noteworthy to mention that possible improvement

of the reliability of the SDQ scores could be reached by a five point Likert response format to improve the reliability of scores [40,60], as well as for dimensional scoring on psychopathology measures [61].

The results of the present study should be interpreted in the light of the following limitations. First, the study is based on adolescents' self-report. As it is well-known, there is a problem in the use of self-report instruments in terms of social desirability and response bias that might be especially important in these age groups. In this sense, the inclusion of clinical indices or behavioural observations could have added objective information. In addition, data from parent and/or teacher would have been useful in order to confirm the information gathered from adolescents and with the aim to test convergent and discriminant validity. Second, the Spanish sample was conformed by different age's group while the others were uniform. This could have implications in the study of the measurement invariance, as the age could be a variable that itself modifies the MI. In addition, this aspect may have implications for the generalizability of our findings to other countries in different age' groups. Further studies could investigate cultural differences as well as national differences and would benefit from including measures of cultural values and beliefs in their assessments. Also, future studies testing the multi-level CFA or IRT should be considered in detecting DIF items and how demographic or economic or cultural variables influence the construct. Future research should continue to advance in the study of measurement invariance of the SDQ dimensions across other nations and/or cultures - in particular using non-Western samples.

On behalf of all authors, the corresponding author states that there is no conflict of interest

References

1. Ortuño J, Fonseca-Pedrero E, Paino M, Aritio-Solana R (2014) Prevalencia de síntomas emocionales y comportamentales en adolescentes españoles [Prevalence of emotional and behavioural symptoms in spanish adolescents]. Rev Psiquiatr Salud Ment 7:121-130

2. Meltzer H, Gatward R, Goodman R, Ford T (2003) Mental health of children and adolescents in Great Britain. Int Rev Psychiatry 15 (1-2):185-187. doi:10.1080/0954026021000046155

3. Gore FM, Bloem PJ, Patton GC, Ferguson J, Joseph V, Coffey C, Sawyer SM, Mathers CD (2011) Global burden of disease in young people aged 10-24 years: a systematic analysis. Lancet 18 (377):2093-2102

4. Erol N, Simsek Z, Oner O, Munir K (2005) Behavioral and Emotional Problems Among Turkish Children at Ages 2 to 3 Years. J Am Acad Child Psy 44 (1):80-87. doi:10.1097/01.chi.0000145234.18056.82

5. Merikangas KR, He JP, Burstein M, Swanson SA, Avenevoli S, Cui L, Benjet C, Georgiades K, Swendsen J (2010) Lifetime prevalence of mental disorders in U.S. adolescents: results from the National Comorbidity Survey Replication--Adolescent Supplement (NCS-A). J Abnorm Child Psychol 49:980-989

6. Kessler RC, Avenevoli S, Costello EJ, Georgiades K, Green JG, Gruber MJ, He JP, Koretz D, McLaughlin KA, Petukhova M, Sampson NA, Zaslavsky AM, Merikangas KR (2012) Prevalence, persistence, and sociodemographic correlates of DSM-IV disorders in the National Comorbidity Survey Replication Adolescent Supplement. Arch Gen Psychiatry 69:372-380

7. Carli V, Hoven CW, Wasserman C, Chiesa F, Guffanti G, Sarchiapone M, Apter A, Balazs J, Brunner R, Corcoran P, Cosman D, Haring C, Iosue M, Kaess M, Kahn JP, Keeley H, Postuvan V, Saiz P, Varnik A, Wasserman D (2014) A newly identified group of adolescents at "invisible" risk for psychopathology and suicidal behavior: findings from the SEYLE study. World Psychiatry 13:78-86

8. Angold A, Messer SC, Stangl D, Farmer E, Costello EJ, Burns BJ (1998) Perceived parental burden and service use for child and adolescent psychiatric disorder. Am J Public Health 88:75-80

9. Ford T, Hamilton H, Meltzer H, Goodman R (2008) Predictors of service use of mental health problems amongs British school children. Child Adolesc Ment Health 13:32-40

10. Goodman R (1997) The Strengths and Difficulties Questionnaire: A Research Note. J Child Psychol Psychiatry 38 (5):581-586

11. Ruchkin V, Jones S, Vermeiren R, Schwab-Stone M (2008) The Strengths and Difficulties Questionnaire: The Self-Report Version in American Urban and Suburban Youth. Psychol Assessment 20 (2):175-182

12. Vostanis P (2006) Strengths and Difficulties Questionnaire: research and clinical applications. Curr Opin Psychiatr 19 (4):367-372. doi: 10.1097/01.yco.0000228755.72366.05

13. Gómez R (2012) Correlated Trait-Correlated Method Minus One Analysis of the Convergent and Discriminant Validities of the Strengths and Difficulties Questionnaire. Assessment XX (X):1-11. doi:10.1177/1073191112457588

14. Klasen H, Woerner W, Wolke D, Meyer R, Overmeyer S, Kaschnitz W, Rothenberger A, Goodman R (2000) Comparing the German versions of the Strengths and Difficulties Questionnaire (SDQ-Deu) and the Child Behavior Checklist. Eur Child Adolesc Psychiatry 9 (4):271-276

15. Muris P, Meesters C, van den Berg F (2003) The Strengths and Difficulties Questionnaire (SDQ). Further evidence for its reliability and validity in a community sample of Dutch children and adolescents. Eur Child Adolesc Psychiatry 12:1-8. doi:10.1007/s00787-003-0298-2

16. Capron C, Therond C, Duyme M (2007) Psychometric Properties of the French Version of the Self-Report and Teacher Strengths and Difficulties Questionnaire (SDQ). Eur J Psychol Assess 23 (2):79-88. doi:10.1027/1015-5759.23.2.79

17. Becker A, Hagenberg N, Roessner N, Woerner W, Rothenberg A (2004) Evaluation of the self-reported SDQ in a clinical setting: Do self-reports tell us more than ratings by adult informants? Eur Child Adolesc Psychiatry y 13 (2):17-24. doi:10.1007/s00787-004-2004-4

18. Koskelainen M, Sourander A, Kaljonen A (2000) The Strenght and Difficulties Questionnaire among Finnish school-aged children and adolescents. Eur Child Adolesc Psychiatry 9 (4):277-284

19. Rønning JA, Helge Handegaard BH, Sourander A, Mørch W-T (2004) The Strengths and Difficulties Self-Report Questionnaire as a screening instrument in Norwegian community samples. Eur Child Adolesc Psychiatry 13:73-82. doi: 10.1007/s00787-004-0356-4

20. Goodman R (2001) Psychometric Properties of the Strengths and Difficulties Questionnaire. J Am Acad Child Psy 40 (11):1337-1345

21. Mellor D (2004) Furthering the Use of the Strengths and Difficulties Questionnaire: Reliability With Younger Child Respondents. Psychol Assessment 16 (4):396-401. doi:doi: 10.1037/1040-3590.16.4.396

22. Mellor D, Stokes M (2007) The Factor Structure of the Strengths and Difficulties Questionnaire. Eur J Psychol Assess 23 (2):105-112. doi: 10.1027/1015-5759.23.2.105

23. Muris P, Maas A (2004) Strengths and difficulties as correlates of attachment style in institutionalized and non-institutionalized children with below-average intellectual abilities. Child Psychiatry Hum Dev 34 (4):317-328. doi:10.1023/B:CHUD.0000020682.55697.4f480164 [pii]

24. Ruchkin V, Koposov R, Schwab-Stone M (2007) The strength and difficulties questionnaire: Scale validation with Russian adolescents. J Clin Psychol 63 (9):861-869

25. Yao S, Zhang C, Zhu X, Jing X, McWhinnie CM, Abela JRZ (2009) Measuring Adolescent Psychopathology: Psychometric Properties of the Self-Report Strengths and Difficulties Questionnaire in a Sample of Chinese Adolescents. J Adolesc Health 45:55-62

26. van de Looij-Jansen PM, Goedhart AW, de Wilde EJ, Treffers PD (2011) Confirmatory factor analysis and factorial invariance analysis of the adolescent self-report Strengths and Difficulties Questionnaire: How important are method effects and minor factors? Br J Clin Psychol 50:127-144

27. He J-P, Burstein M, Schmitz A (2012) The Strengths and Difficulties Questionnaire (SDQ): the Factor Structure and Scale Validation in U.S. Adolescents. J Abnorm Child Psychol

28. Svedin CG, Priebe G (2008) The Strengths and Difficulties Questionnaire as a screening instrument in a community sample of high school seniors in Sweden. Nord J Psychiatr 62 (3):225-232. doi: 10.1080/08039480801984032

29. Van Roy B, Veenstra M, Clench-Aas J (2008) Construct validity of the five-factor Strengths and Difficulties Questionnaire (SDQ) in pre-, early, and late adolescence. J Child Psychol Psychiatry 49 (12):1304-1312. doi:10.1111/j.1469-7610.2008.01942.x

30. Di Riso D, Salcuni S, Chessa D, Raudino A, Lis A, Altoè G (2010) The Strengths and Difficulties Questionnaire (SDQ). Early evidence of its reliability and validity in a community sample of Italian children. Pers Individ Diff 49:570-575

31. Dickey WC, Blumberg SJ (2004) Revisiting the Factor Structure of the Strengths and Difficulties Questionnaire: United States, 2001. J Am Acad Child Psy 43 (9):1159-1167. doi:10.1097/01.chi.0000132808.36708.a9

32. Percy A, McCrystal P, Higgins K (2008) Confirmatory Factor Analysis of the Adolescent Self-Report Strengths and Difficulties Questionnaire. Eur J Psychol Assess 24 (1):43-48. doi:10.1027/1015-5759.24.1.43

33. Goodman A, Lamping DL, Ploubidis GB (2010) When to Use Broader Internalising and Externalising Subscales Instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British Parents, Teachers and Children. J Abnorm Child Psychol 38:1179-1191. doi:10.1007/s10802-010-9434-x

34. Caci H, Morin AJ, Tran A (in press) Investigation of a bifactor model of the Strengths and Difficulties Questionnaire. Eur Child Adolesc Psychiatry

35. Goodman A-, Heiervang E, Fleitlich-Bilyk B, Alyahri A, Patel V, Mullick MS, Slobodskaya H, Dos Santos DN, Goodman R (2012) Cross-national differences in questionnaires do not necessarily reflect comparable differences in disorder prevalence. Soc Psychiatry Psychiatr Epidemiol 47 (8):1321-1331. doi: 10.1007/s00127-011-0440-2

36. Milfont TL, Fisher R (2010) Testing measurement invariance across groups: applications for cross-cultural research. Int J Meth Psych Res 3:111-121

37. Byrne B (2012) Structural Equation Modeling with Mplus: Basic concepts, applications, and programming. Routledge Taylor & Francis Group, New York

38. Essau CA, Olaya B, Anastassiou-Hadjicharalambous X, Pauli G, Gilvarry C, Bray D, O'Callaghan J, Ollendick TH (2012) Psychometric properties of the Strengths and Difficulties Questionnaire from five European countries. Int J Meth Psych Res 21 (3):232-245. doi:10.1002/mpr.1364

39. Stevanovic D, Urbán R, Atilola O, Vostanis P, Singh Balhara YP, M. Avicenna M, Kandemir H, Knez R, Franic T, Petrov P (2014) Does the Strengths and Difficulties Questionnaire – self report yield invariant measurements across different nations? Data from the International Child Mental Health Study Group. Epidemiol Psychiatr Sci 30:1-12. doi: 10.1017/S2045796014000201

40. Ortuño-Sierra J, Fonseca-Pedrero E, Paino M, Sastre i RIba S, Muñiz J (2015) Screening mental health problems during adolescence: Psychometric properties of the Spanish version of the Strengths and Difficulties Questionnaire. J Adolesc

41. Fonseca-Pedrero E, Paino M, Lemos-Girádez S, Muñiz J (2011) Prevalencia de la sintomatología emocional y comportamental en adolescentes españoles a través del Strengths and Difficulties Questionnaire (SDQ). Rev. Psicopatol. Psicol. Clín 16:15-25

42. Schumann G, Loth E, Banaschewski T, Barbot A, Barker G, Büchel C, Conrod PJ, Dalley JW, Flor H, Gallinat J, Garavan H, Heinz A, Itterman B, Lathrop M, Mallik C, Mann K, Martinot JL, Paus T, Poline JB, Robbins TW, Rietschel M, Reed L, Smolka M, Spanagel R, Speiser C, Stephens DN, Ströhle A, Struve M, IMAGEN consortium (2010) The IMAGEN study: reinforcement-related

behaviour in normal brain function and psychopathology. Mol Psychiatry 15 (12):1128-1139. doi:10.1038/mp.2010.4.

43. Muthén LK, Muthén BO (1998-2012) Mplus User's Guide. Seventh Edition. Muthén & Muthén, Los Angeles

44. Brown TA (2006) Confirmatory Factor Analysis for Applied Research. Guilford Press, New York

45. Hu LT, Bentler PM (1999) Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Modeling 6 (1):1-55

46. Heene M, Hilbert S, Freudenthaler HH, Bühner M (2012) Sensitivity of SEM Fit Indexes With Respect to Violations of Uncorrelated Errors. Struct Equ Modeling 19:36-50

47. Byrne BM (2008) Testing for multigroup equivalence of a measuring instrument: A walk through the process. Psicothema 20:872-882

48. Muthén BO, Asparouhov T (2002) Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Mplus Web Note No. 4.http://www.statmodel.com/mplus/examples/webnote.html

49. Byrne BM, Stewart SM (2006) The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. Struct Equ Modeling 13:287-321

50. Meredith W (1993) Measurement invariance, factor analysis and factorial invariance. Psychometrika 58:525-543

51. Horn JL, McArdle JJ (1992) A practical and theoretical guide to measurement invariance in aging research. Exper Ag Research 18 (3-4):117-144

52. Cheung GW, Rensvold RB (2002) Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equ Modeling 9:233-255. doi:doi:10.1207/S15328007SEM0902_5

53. Byrne BM, Shavelson RJ, Muthén B (1989) Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychol Bull 105:456-466

54. Statistical Package for the Social Sciences (2006) SPSS Base 15.0 User's Guide. SPSS Inc, Chicago, IL

55. Marsh HW, Hau KT, Wen Z (2004) In search of golden rules: Comment on hypothesis testing approaches to setting cut-off values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Struc Equ Modeling 11:320-341

56. Koskelainen M, Sourander A, Vauras M (2001) Self-reported strengths and difficulties in a community sample of Finnish adolescents. Eur Child Adolesc Psychiatry 10:180-185

57. Lerner RM, Galambos NL (1998) Adolescent development: challenges and opportunities for research, programs, and policies. Annu Rev Psychol 49:413-446

58. Ivanova MY, Achenbach TM, Rescorla LA, Dumenci L, Almqvist F, Bilenberg N, Bird H, Broberg AG, Dobrean A, Döpfner M, Erol N, Forns M, Hannesdottir H, Kanbayashi Y, Lambert MC, Leung P, Minaei A, Mulatu MS, Novik T, Oh KJ, Roussos A, Sawyer M, Simsek Z, Steinhausen HC, Weintraub S, Winkler Metzke C, Wolanczyk T, Zilber N, Zukauskiene R, Verhulst FC (2007) The generalizability of the youth self-report syndrome structure in 23 societies. J Consult Clin Psych 75:729-738

59. Muris P, Meesters C, Eijkelenboom A, Vincken M (2004) The self-report version of the Strengths and Difficulties Questionnaire: Its psychometric properties in 8- to 13- year-old non-clinical children. Brit J Clin Psychol 43 (4):437-448. doi:10.1348/0144665042388982

60. Zumbo BM, Gadermann AM, Zeisser C (2007) Ordinal versions of coefficients alpha and theta for Likert rating scales. J Mod Appl Stat Methods 6:21-29

61. Markon KE, Chmielewski M, Miller CJ (2011) The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. Psychol Bull 137:856-879

Table 1
Goodness-of-fit indices of the models tested in the confirmatory factor analysis

Models	$\chi 2$	*df*	CFI	TLI	RMSEA (CI 90%)	WRMR
(a) Baseline Five-factor						
Spain	1010.45	265	.823	.794	.067 (.063-.074)	1.75
England	895.15	265	.842	.823	.064 (.056-.067)	1.65
Ireland	453.17	265	.807	.792	.068 (.062-.074)	1.17
Germany	804.19	265	.872	.854	.045 (.041-.053)	1.51
France	406.75	265	.889	.883	.043 (.040-.054)	1.10
Total	2371.74	265	.859	.842	.052 (.048-.061)	2.56
(b) Five-factor with CE added						
Spain	653.03	260	.904	.891	.049 (.041-.054)	1.36
England	679.85	260	.892	.872	.054 (.049-.061)	1.41
Ireland	421.36	260	.842	.821	.052 (.047-.061)	1,11
Germany	656.50	260	.909	.888	.049 (.041-.052)	1.35
France	341.64	260	.938	.932	.038 (.031-.042)	.97
Total	1527.70	260	.912	.903	.045 (.041-.052)	2.04
(c) Five-factor with CE and reverse-worded items						
Spain	545.88	255	.934	.924	.045 (.038-.049)	1.21
England	568.03	255	.923	.913	.046 (.041-.052)	1.24
Ireland	361.38	255	.901	.878	.049 (.042-.053)	.99
Germany	610.47	255	.914	.901	.048 (.041-.052)	1.27
France	323.56	255	.947	.941	.039 (.030-.041)	.93
Total	1187.99	255	.934	.932	.044 (.037-.049)	1.76
(d) Three-factor						
Spain	1521.96	272	.692	.659	.071 (.067-.078)	2.21
England	1.125.22	272	.784	.759	.066 (.061-.072)	1.91
Ireland	550.694	272	.723	.689	.079 (.069-.081)	1.35
Germany	1266.40	272	.764	.742	.063 (.061-.073)	1.98
France	518.36	272	.818	.802	.060 (.056-.068)	1.33
Total	3694.36	272	.768	.743	.068 (.062-.073)	3.30
(e) Three-factor with CE added						
Spain	954.05	267	.826	.814	.049 (.042-.054)	1.70
England	857.89	267	.854	.832	.047 (.043-.055)	1.64
Ireland	508.17	267	.757	.732	.064 (.061-.073)	1.28
Germany	1079.35	267	.804	.783	.054 (.047-.059)	1.82
France	436.80	267	.873	.858	.046 (.041-.052)	1.18
Total	2572.72	267	.842	.819	.049 (.043-.054)	2.72
(f) Three-factor with CE and reversed-worded items						
Spain	752.90	262	.883	.857	.047 (.042-.051)	1.47
England	734.46	262	.884	.856	.046 (.041-.050)	1.47
Ireland	420.69	262	.837	.824	.046 (.031-.054)	1.11
Germany	949.68	262	.832	.812	.049 (.041-.054)	1.67
France	383.88	262	.914	.903	.047 (.041-.053)	1.07

Total	2046.40	262	.884	.856	.047 (.041-.052)	2.38
(g) Five-factor and two second-order factor						
Spain	1150.40	268	.778	.762	.062(.058-.074)	1.89
England	922.50	268	.832	.812	.064 (.057-.071)	1.71
Ireland	490.80	268	.779	.753	.062 (.056-.073)	1.25
Germany	895.32	268	.854	.832	.053 (.047-.062)	1.63
France	419.79	268	.892	.871	.052 (.044-.059)	1.14
Total	2699.40	268	.836	.821	.061 (.048-.069)	2.78
(h) Five-factor and two second-order factor with CE						
Spain	800.120	263	.868	.849	.053 (.046-.061)	1.537
England	732.316	263	.879	.862	.052 (.046-.061)	1.493
Ireland	469.837	263	.793	.764	.061 (.054-.071)	1.207
Germany	766.583	263	.878	.861	.042 (.038-.052)	1.489
France	356.783	263	.930	.920	.038 (.031-.044)	1.010
Total	1933.338	263	.887	.871	.048 (.040-.052)	2.333
(i) Five-factor and two second-order factor with CE and reverse-worded items						
Spain	608.795	258	.914	.900	.042 (.037-.055)	1.304
England	571.569	258	.919	.906	.043 (.361-.052)	1.271
Ireland	382.452	258	.876	.855	.051 (.044-.058)	1.040
Germany	641.278	258	.908	.892	.043 (.037-.052)	1.338
France	332.373	258	.945	.936	.034 (.028-.042)	.950
Total	1313.071	258	.928	.917	.038 (.031-.047)	1.885

Note. χ^2 = Chi square; *df* = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; WRMR= Weighted Root Mean Square Residual. CE = Correlated Errors: 2-10, 25-15, 15-16, 19-18, 20-23.

Table 2

Goodness-of-fit indices for measurement invariance of the SDQ across European countries

	χ^2	*df*	CFI	TLI	RMSEA (90% CI)	WRMR	ΔCFI	Model Comparison
Five-Factor with five CE								
Configural Invariance (1)	2642.712	1300	.903	.892	.041 (.04-.05)	2.79		
Strong factorial invariance (2)	3313.931	1460	.863	.864	.046 (.04-.06)	3.30	0.040	2 vs 1
Strong partial factorial invariance* (3)	3020.123	1457	.895	.897	.043 (.04-.05)	3.115	0.008	3 vs 1

Note. χ^2 = Chi square; *df* = degrees of freedom; CFI = Confirmatory Factor Index; TLI= Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; WRMR = Weighted Root Mean Square Residual; CI = Confidence Interval; ΔCFI = Change in Comparative Fit Index; CE = Correlated Errors; *(freeing factor loadings and thresholds of items: 2, 15, 20, 21-England-; 1, 6, 7, 15, 16, 17, 21, 22, 23-Germany-; 21-France-)

Table 3

Standardized factor loadings for the strong partial measurement invariance model

	Spain	England	Ireland	Germany	France
Items	Loadings (R^2)	Loadings (R^2)	Loadings (R^2)	Loadings (R^2)	Loadings (R^2)
Emotional Problems					
3	**.47 (.22)**	**.47 (.22)**	**.46 (.21)**	**.46 (.21)**	**.53 (.28)**
8	**.68 (.47)**	**.63 (.39)**	**.60 (.37)**	**.66 (.44)**	**.76 (.57)**
13	**.72 (.51)**	**.79 (.62)**	**.97 (.93)**	**.84 (.70)**	**.80 (64)**
16	.59 (.35)	.56 (.32)	.52 (.27)	.52 (. 27)	.63 (.39)
24	**.60 (.36)**	**.54 (.29)**	**.50 (.25)**	**.65 (.43)**	**.68 (.46)**
Conduct Problems					
5	**.46 (.21)**	**.65 (.43)**	**.67 (.45)**	**.48 (.23)**	**.51 (.36)**
7	.45 (.20)	.53 (.28)	.59 (.34)	.66 (. 16)	.42 (.17)
12	**.58 (.33)**	**.65 (.43)**	**.55 (.30)**	**.56 (.31)**	**.76 (.58)**
18	**.46 (.22)**	**.57 (.33)**	**.58 (.34)**	**.70 (.22)**	**.55 (.31)**
22	.48 (24)	.54 (.30)	.56 (.31)	.50 (.18)	.49 (.24)
Peer Problems					
6	.66 (.44)	.59 (.34)	.46 (.21)	.46 (.21)	.58 (.33)
11	**.73 (.39)**	**.57 (.32)**	**.46 (.21)**	**.66 (.43)**	**.85 (.73)**
14	**.67 (.45)**	**.58 (.33)**	**.58 (.34)**	**.56 (.31)**	**.65 (.42)**
19	**.52 (.27)**	**.66 (.43)**	**.73 (.54)**	**.70 (.49)**	**.79 (.62)**
23	.41 (.17)	.41 (.17)	.49 (.24)	.50 (.25)	.46 (.21)
Hyperactivity					
2	.43 (.19)	.57 (.32)	.56 (.31)	.54 (.29)	.54 (.29)
10	**.39 (.15)**	**.54 (.30)**	**.57 (.33)**	**.49 (.24)**	**.57 (.33)**
15	.57 (.33)	.75 (.57)	.63 (.40)	.72 (.52)	.70 (.50)
21	.49 (.24)	.64 (.41)	.55 (.30)	.61 (.37)	.59 (.34)
25	**.47 (.22)**	**.70 (.49)**	**.62 (.38)**	**.67 (.45)**	**.55 (.30)**
Prosocial					
1	.72 (.52)	.74 (.55)	.77 (.60)	.71 (.50)	.77 (.42)
4	**.49 (.24)**	**.57 (.33)**	**.61 (.37)**	**.48 (.23)**	**.61 (.17)**
9	**.58 (.34)**	**.75 (.56)**	**.75 (.56)**	**.72 (.51)**	**.75 (.47)**
17	.62 (.39)	.66 (.44)	.60 (.36)	.61 (.38)	.60 (.25)
20	.51 (.26)	.53 (.28)	.47 (.22)	.51 (.26)	.47 (.27)

Note. All standardized factorial loadings estimated were statistically significant ($p < .01$); R^2 = Proportion of explained variance; factor loadings equivalent and non-equivalent across countries differentiated in bold and normal font, respectively

Table 4

Ordinal alpha coefficients for the total sample and for each country in the five-factor model and the final model with reverse items included

	Spain (n = 848)	England (n = 626)	Ireland (n = 227)	Germany (n = 1050)	France (n = 260)	Total (n = 3012)
Emotional	.78	.70	.72	.73	.74	.78
Conduct	.68	.67	.66	.62	.61	.69
Peer Problems	.75	.62	.61	.73	.64	.70
Hyperactivity	.70	.79	.74	.74	.83	.76
Prosocial	.74	.77	.75	.77	.83	.76
Total Difficulties	.84	.83	.81	.75	.85	.83

Table 5

Mean and standard deviation of the SDQ Total Difficulties scores and the subscales across the five countries

	Spain	England	Ireland	Germany	France	Total
	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
Emotional	3.19 (2.27)	2.87 (2.07)	2.27 (2.04)	2.57 (2.02)	2.48 (2.19)	2.81 (2.14)
Conduct	2.40 (1.60)	2.25 (1.70)	2.28 (1.68)	1.91 (1.33)	2.47 (1.66)	2.20 (1.56)
Peer Problems	1.90 (1.81)	1.86 (1.64)	1.49 (1.50)	2.00 (1.60)	1.34 (1.50)	1.85 (1.67)
Hyperactivity	4.46 (2.16)	4.49 (2.25)	4.28 (2.24)	3.75 (2.00)	3.81 (2.19)	4.15 (2.16)
Prosocial	8.17 (1.66)	7.54 (1.72)	7.52 (1.76)	7.82 (1.63)	7.59 (1.64)	7.82 (1.69)
Total Difficulties	11.94 (5.21)	11.47 (5.07)	10.72 (5.05)	10.24 (4.31)	10.11 (4.85)	11.00 (4.90)